What is Data Science ?*
Fundamental Concepts and a Heuristic Example

Chikio Hayashi

The Institute of Statistical Mathematics
Sakuragaoka, Birijian 304
15-8 Sakuragaoka, Shibuya-ku
Tokyo 150, Japan

Summary: Data Science is not only a synthetic concept to unify statistics, data analysis and their related methods but also comprises its results. It includes three phases, design for data, collection of data, and analysis on data. Fundamental concepts and various methods based on it are discussed with a heuristic example.

1. Introduction:

Statistics and data analysis have developed in their realms separately and contributed to the development of science, showing their unique properties. The ideas and various methods of statistics were very useful, well known and solved many problems. Mathematical statistics succeeded it and developed new frontiers with the idea of statistical inference. Thus the application of these view points brought us many useful results.
However, the development of mathematical statistics, has devoted itself only to the problems of statistical inference, an apparent rise of precision of statistical models, and to the pursuit of exactness and mathematical refinement, so mathematical statistics have been prone to be removed from reality.
On the other hand, the method of data analysis has developed in the fields disregarded by mathematical statistics and has given useful results to solve complicated problems based on mathematico-statistical methods (which are not always based on statistical inference but rather are descriptive). Some results are found in the references.
In the development of data analysis, the following tendency is often found, that is to say, data analysts have come to manipulate or handle only existing data without taking into consideration both the quality of data and the meaning of data, to cope with the methodological problem based on unrealistic artificial data with simple structure, to make efforts only for the refinement of convenient and serviceable computer software and to imitate popular ideas of mathematical statistics without considering the essential meaning.
As this differentiation proceeds with specialization, the innovation of useful methods of statistics and data analysis seem to disappear and signs of stagnation appear. The reason is that the essential aim of analysis of phenomena with data has been forgotten. For extensive and profound development of intrinsically useful methods of statistics and data analysis beyond the present state, the unification of statistics and data analysis is necessary. For this purpose, the construction of a new point of view or a new paradigm is a crucial problem. So, I will present "Data Science" as a new concept.

---

* The roundtable discussion "Perspectives in classification and the Future of IFCS" was held at the last Conference under the chairmanship of Professor H. -H. Bock. In this panel discussion, I used the phrase 'Data Science'. There was a question, "What is 'Data Science'? " I briefly answered it. This is the starting point of the present paper.

C. Hayashi et al. (eds.), *Data Science, Classification, and Related Methods*

2. Fundamental Concepts of Data Science

Data Science is not only a synthetic concept to unify statistics, data analysis and their related methods, but also comprises its results. Data Science intends to analyze and understand actual phenomena with "data". In other words, the aim of data science is to reveal the features or the hidden structure of complicated natural, human and social phenomena with data from a different point of view from the established or traditional theory and method. This point of view implies multidimensional, dynamic and flexible ways of thinking.
Data Science consists of three phases : design for data, collection of data and analysis on data. It is important that the three phases are treated with the concept of unification based on the fundamental philosophy of science explained below. In these phases the methods which are fitted for the object and are valid, must be studied with a good perspective. The strategy for research in Data Science through three phases is summarized in Fig. 1.

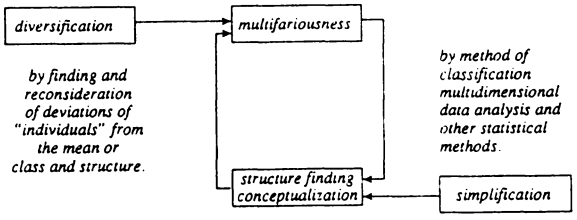
Data Science

Design for Data

Collection of Data

Analysis on Data

The leading idea of data treatment in the research process is:



Dynamics of Both Diversification

and Conceptualization or Simplification

Fig.1 Strategy for Research

Generally speaking, phenomena are multifarious. First, these phenomena are formulated and the planning of a survey or experiment is completed, based on the ideas of Data Science (phase of design for data). Thus phenomena are expressed as multidimensional and, frequently, time-series data. The characteristics or properties of the data are necessarily made clear (phase of collection of data). The obtained data are too complicated to draw a clear conclusion. So, by methods of classification and multidimensional data analysis, and other mathematico-statistical methods, the data structure is revealed. In other words, simplification and conceptualization are carried out. However, this information generally turns out to be incomplete and unsatisfactory

even though the structure finding was realized. At this stage, by finding and reconsidering the deviation of "individuals", which gives a vivid account of the roughness of conceptualization or simplification, from the mean values or class-belonging (classification) and structure, diversification of data is made. Based on this multifariousness, structure finding or conceptualization is attained, in an advanced sense, in the progressive stages. Such a circular movement of research then continues. Dynamic movement of both simplification or conceptualization and diversification begins in turn. Further, having been able to solve a problem, it is expected to discover another new problem to be solved in an advanced sense. The developmental process, in phase, design --->collection –> analysis --->design --->collection --->analysis --->design –> collection ---> analysis ---> design ---> ... and the dynamic process mentioned above, that is to say, progress and regress, are indispensable in Data Science. This shows that the methodology of Data Science develops, as it were, in the ascending-spiral-process and research proceeds as seen in spiral stairs. The main points is schematically depicted in Fig. 1.
Thus we can say that data science comprises not only the results themselves of theory and method but also all methodological results related to various processes which are necessary to work out the results mentioned above. The former is called "hard results" and the latter is called "soft results". Data Science includes simultaneously hard and soft results. It goes without saying that a useful solution emerges in coping with the complicated problem in question by the use of Data Science. It is repeatedly emphasized that the coherent idea through all items shown in Fig. 1 flows in Data Science for the purpose of analysis of phenomena with data.

3. Content of Data Science

Some concrete examples in social and medical surveys for the three phases are shown below. Before everything, it is stressed that the relevant methods are always treated with validity.

3.1 How to Design

The theory and method concerning this phase are next considered. Particularly, theoretical and systematic construction of a questionnaire is a very important problem. The problems in this phase are frequently solved using various kinds of methods of data analysis. For example,

Sampling survey methods,
Design of experiment,
Evaluation of bias in quota sampling
New systematic idea of survey planning for the solution of difficult and complicated problems,
Construction of questionnaire,
    Theory-driven (which is an extension of hypothesis testing), Guttman's Facet Theory,
    Data-driven (exploratory approach), Hayashi's Cultural Link Analysis in comparative study,
Utilization of various types of questions, for example, dynamic use of closed and open-ended questions,
Use of various projective methods,
Design for evaluation of data quality and data characteristics,
Randomized response method,
Problems of translation in international comparative study,
and etc..

3.2 How to Collect Data

Collection of data is not only a problem of practice, but must be theoretically and concretely studied. The problems in this phase can not be solved without any information of design for data and any use of data analysis.

Evaluation of survey bias and evaluation of experimental bias including question bias, interview bias, interviewer bias, observation bias, etc.
Evaluation of non-response error,
Evaluation of measurement error,
Evaluation of response error, inevitably variable response data, for example, live data,
Method of diminution of the relevant bias and error,
and etc..

3.3 How to Analyze Data

The problems in this phase are, of course, closely related to the previous two phases. The main point is to obtain useful and instrumental information without any distortion or with validity. For this purpose, clear and lucid methods of analysis without unnecessary mathematical conditions only for model building and a too sophisticated style are desirable. For example,

Various methods of scaling, quantification methods, correspondence analysis (analyse des données), multidimensional scaling, exploratory data analysis, categorical data analysis and various methods of classification and clustering,
Useful data analysis suitable for the purpose,
Useful coding of questions and their synthesis,
Valid analysis of data including various errors,
Evaluation of data quality and data analysis depending on data quality,
Analysis on probabilistic response,
Exploratory approach by data analysis,
Method of simultaneous realization of classification and structure finding,
Treatment of open answers in an open ended question for example, exploratory approach for coding or automatic processing of textual data,
Probabilistic approach,
Computer experiments,
and etc..

These three phases must be synthetically treated or taken into consideration with the consistent idea in order to understand phenomena. This is the fundamental concept of Data Science. Of course, each subject will be studied separately. However, each subject must be studied in the context of Data Science. This idea will lead to the development of statistics and data analysis in a new direction. Thus the stand point of them highten and a new horizon will appear as innovative method and theory are created in three phases.

4. A Heuristic Example

As an example of the data-scientific approach, we now explain our national character survey in interchronological and international perspectives.

4.1 Fundamental Scheme of Study

What is national character ? I define it operationally as collective character on belief systems, the way of thinking and emotional attitudes, feelings or sentiments. By a survey of individuals, we can find individual response patterns on the items mentioned below.

Thus, we know that individuals have various response patterns. These are integrated in a collective through mutual and social communications in so far as individuals live in a society. This is collective character or national character (in some cases ethnic character) which is formed beyond individuals. In this situation, some principles emerge in the social environment. Receiving impacts from the exterior, social norm, customs system, paradigm, education, contemporary thought and arts, religious feelings, future course of philosophy and science, etc. are formed, as a "cultural climate" is created. Individuals are influenced by this cultural climate: the strongest influence is upon the response pattern in general social items, the second upon that in national character items and the weakest upon that in basic human feelings items. Such a perpetual circular movement continues. It is our aim to represent the collective character in terms of Data Science.
Our point of view of research is not hypothesis testing (theory-driven) but to put the emphasis on an exploratory approach (data-driven).

4.2 Time Series Data (Interchronological Approach) in Japan

First of all, we define the universe and population of the Japanese. A nation-wide sample survey is done for a sample by stratified three stage random sampling and by face to face interviewing using the same questionnaire, the contents of which cover the items shown below.

1) Fundamental Attributes, 2) Religion, 3) Family,
4) Social Life, 5) Interpersonal Relations,
6) Politics, 7) Individual Attitude toward Other Unclassified Social Issues

The outline of our survey is shown in Fig. 2.

Nation Wide Sample Survey by
Stratified Three Stage Random Sampling

Sample Spot 200 - 300
Sample Size 2000 - 4000
by face-to-face interviewing

Survey	Symbol	Year
1st Survey	I	1953
2nd Survey	II	1958
3rd Survey	III	1963
4th Survey	IV	1968
5th Survey	V	1973
6th Survey	VI	1978
7th Survey	VII	1983
8th Survey	VIII	1988
9th Survey	IX	1993
		every 5 years

[Research Committee on the Study of Japanese National Character of the Institute of Statistical Mathematics, Tokyo]

Fig. 2 Survey Design

The analysis from such time series data makes clear both enduring and changing aspects. The next step is a comparative study of national character.

4.3 Comparative Study (International Approach)

In a comparative study, the following points are indispensable,

1. How to secure comparability in a scientific sense
   - Design
     - Sample
     - Selection of questions and construction of questionnaire
     - Translation*
2. Clarification of particularity and universality (community) or speciality and generality
3. By common logic and scientific methods for easy (international understanding)

---

* back translation, retranslation, confirmation by free question and answers, etc.

Here, in a comparative study, we present a new idea for questionnaire construction and selection of nations to be compared. This is Cultural Link Analysis (CLA in abbreviation) --Hayashi et al. 1986, 1992-- which belongs to a similar genre to Guttman's Facet Theory, (Guttman (1994)), and reveals a relational structure of collective characters of peoples in different cultural spheres (nations or ethnic groups).

a. A spatial link inherent in the selection of the subject culture or society. The connections seen in such selection may be considered along the dimensions of social environment, culture and ethnic or national characteristics.

b. An item structure link inherent in the commonness and differences in item response patterns within and across different cultures.

c. A temporal link inherent in longitudinal analysis.

An example of **a.** is shown in Fig.3.

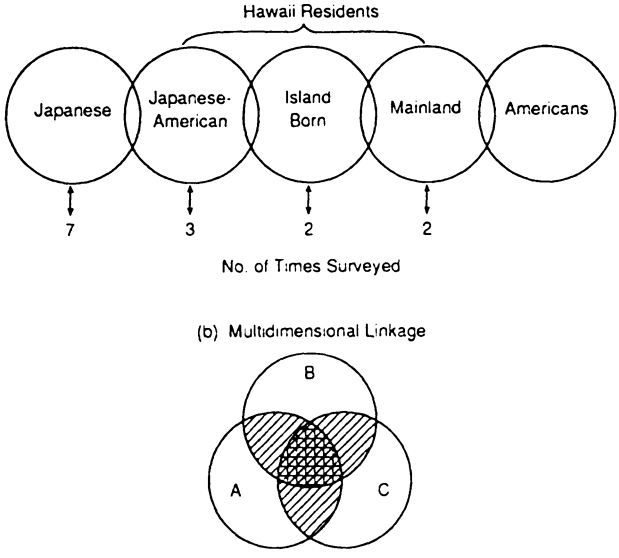


Fig.3 Cultural link survey design: selection of groups

As an example of **b**. concerning questionnaire construction, the idea is explained in Fig.4.

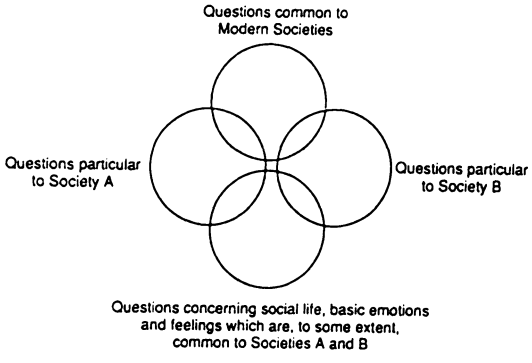


Fig.4 Cultural Link Survey Design: Selection of Questions

As for **c**. time series surveys in various nations or ethnic groups and their comparison are informative.
Our international comparative surveys, which consist of Americans in North America, English in UK., French in France, Germans in the past West-Germany, Dutch in the Netherlands, Italians in Italy, Japanese in Japan, Japanese-Americans in Hawaii, Japanese-Brazilians in Brazil, are described in Fig. 5 and the conjecture of link scheme is depicted in Fig.6.

1971	Japanese Americans in Hawaii (434)		
1978	Honolulu residents including JA (751)	Americans in North America (1571)	
1983	Same above (807)	----------	
1987	----------	----------	English (1043) Germans (1000) French (1013)
1988	Same above (499)	Americans in North America (1563)	
1992	Japanese Brazilians in Brazil (492)		Italians (1048)
1993			Dutch (1083)

------Nation-Wide Sampling------
( ) Sample Size

Fig.5. Comparative Surveys By Cultural Link Analysis

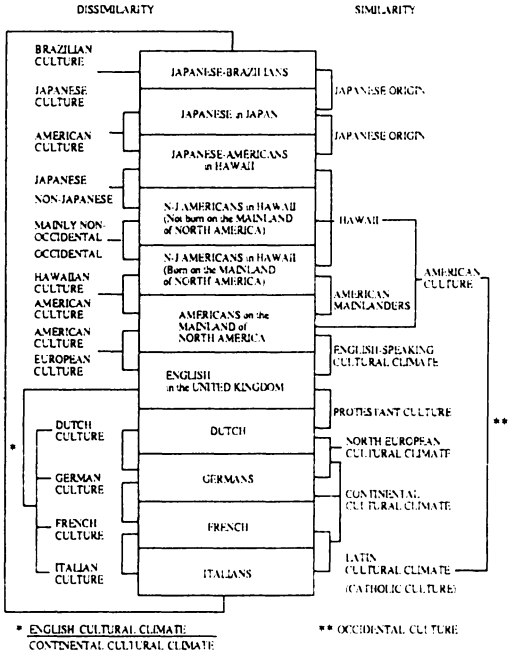


Fig.6 Chain in Our Study

Further Remarks : It brings us very important information to include people of Japanese origin, who settled down in foreign countries, as a linkage in order to explore and reveal the characteristics of Japanese national character.
The attitude, in which Japanese Americans are between Japanese and Americans in response distributions or data structure and, what is more, Japanese Brazilians are between Japanese and Portuguese, French and Italian in response distributions or data structure, is defined as J-attitude. The existence of J-attitude implies that J-attitude is a characteristic of Japanese national character. In other words, it may be said that J-attitude remains somewhat in Japanese Americans and Japanese Brazilians even though the tendency in them is not so strong as in the Japanese. This fact suggests that J-attitude is a Japanese characteristic. [Hayashi (1995), Hayashi C. and F. (1995)] So, it is meaningful to include people of Japanese origin in a comparative survey. However, it goes without saying that the characteristics of the Japanese are found according to the items shown in Fig.8, even though they are not J-attitude.

4.4 National Character in Statistical Terms

It is our aim to make clear and depict the following points by well-designed comparative surveys and their data analysis, i.e. "quantitative and data scientific" methods,

"difference in some points and commonness or similarity in other points"
or
"particularity in some points and universality in other points"

Since such a way of research is based on universal logic, people even in different cultural spheres can understand the results of the analysis.

Mainly considering the view of Japanese national character itself, we can summarize our study as in Fig.7. In contrast, mainly considering the comparison of national character in different cultural spheres, we can summarize our study as in Fig. 8.

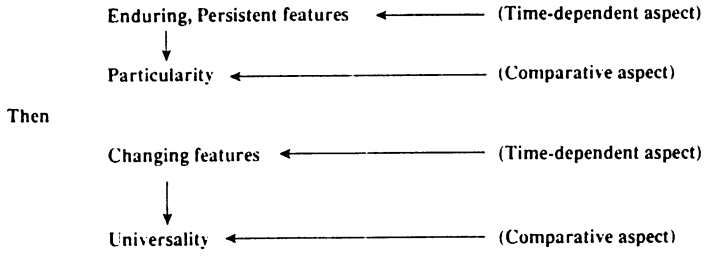


Fig.7 Japanese National Character

From these two kinds of surveys, surveys both in time and space, that is to say, continuing surveys and comparative surveys, we can define national character in statistical terms corresponding to various levels. See Fig. 8.

	Temporally Stable Consistent	Particular Characteristic compared with others
1. Majority Opinion	O O no datum	O* or O indifferent O or O*
2. Opinion Distributions	O no datum	O* or O O or O*
3. Opinion Distributions by Various Breakdowns (for example, gender, age education, rural-urban)	O no datum	O* or O O or O*
4. Changing Patterns of Opinion Distributions and Opinion Structure (including those by breakdowns, and, for example, age-cohort analysis based on time series of opinion distributions)	X	O
5-1 Opinion Structure	O or systematic change	X
5-2 Comparison of Opinion Structures		
i Existence of the Same Unidimensional Scale	O or no datum	by comparison of the scale value of nation O* or O
ii Same Structure of Opinions (more than 2 dimensional structure)	O or no datum	by comparison of the position of nation O or O*
iii Different Structure of Opinions	O or no datum	by comparison of the position of nation based on the similarity or dissimilarity analysis of structure O or O*

Fig. 8 Statistical Definition of National Character
---on various levels---

Here, majority opinion is defined as not only that supported by more than 2/3 of the individuals in the total but also that supported by more than 2/3 of the individuals in each breakdown in sex, age and education. In Fig. 8, O marks mean existence of the item, O* marks mean existence of temporally stable data and "no datum" means non existence of temporally stable evidence but existence of cross-section data. For example, as for 2. Opinion Distributions, the first line means a definition on the highest level, i.e. the opinion distribution is not only temporally stable but also particular or characteristic compared with those in different nations or ethnic groups and the second line means a definition on a lower level, i.e. temporally stable data do not exist but it is particular or characteristic compared with those in different nations or ethnic groups, in which temporally stable evidence occasionally exists. X marks mean there is no logical meaning.

4.5 Cross-Societal Surveys and Classification of Nations --Realization of Cultural Link--

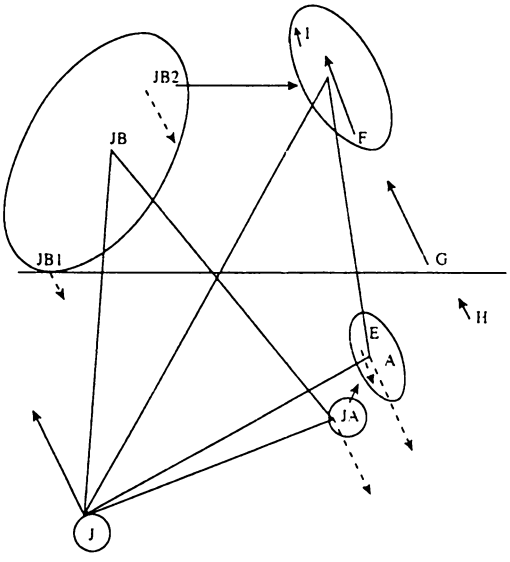
One example of data analysis of comparative surveys will be shown as below, with the following groups being taken up: Japanese, Americans, English, French, Germans, Dutch, Italians, Japanese-Americans and Japanese-Brazilians.
For example, let the opinion distribution be given in each group. Here, only one key answer category is taken up in each question item. If the number of questions is R, the number of the answer category taken up is r. Here, all questions are used except for the items of personal characteristics, for example sex, age, education, etc. We calculate the similarity index d_{ij} between i-nation and j-nation, as below.

$$d_{ij} = 1/R \sum_{r}^{R} | P_{ir} - P_{jr} |$$

where P_{ir} is the percentage of i-nation on the only one key answer category of the r-th question.
d_{ij} is a fuzzy measure of difference between i and j.

Thus we have a similarity matrix between i and j. Based on this fuzzy similarity matrix, a method of multidimensional data analysis, MDA-OR (Minimum Dimension Analysis Ordered class belonging) [Hayashi 1974, 1976], which is one kind of so-called multidimensional scaling MDS, is applied for graphical representation of groups. The quite similar configuration of groups is obtained by quantification method III or Correspondence analysis using the matrix of d's directly. The result is shown in Fig.9. This is a simple graphical summarization of the similarity relations. The degree of similarity is revealed as the distance in Euclidean space. Roughly speaking, consider that the distance corresponds to the similarity and the configuration gives a reasonable summarization of linked similarities. Here, the triangular relation mentioned above has been revealed.

The arrow means the direction of the value in the third axis in Fig.9. A line means plus direction while a dotted line means minus direction in the third dimension.
If French and Italians are deleted and the same analysis is done, Fig. 10 is obtained.
JB is found as a pole instead of French and Italians.



A : Americans
E : English
F : French
G : Germans
H : Dutch
I : Italians
J : Japanese
JA : Japanese-Americans in Hawaii
JB : Japanese-Brazilians in Brazil
JB1: First generation of JB
JB2: Second and Third generations of JB

Fig. 9 Configurations of Nations

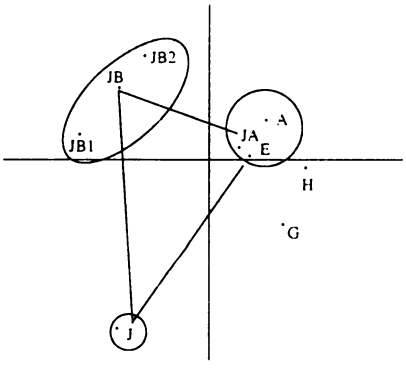


Fig.10 Configurations of Nations (Deleting French and Italians)

The similarity between Japanese-Brazilians and French and Italians is very interesting. The positions of Japanese-Americans and Japanese-Brazilians are to be noted with J-attitude mentioned previously, the former being as a linkage between Japanese and Americans, the latter being as a linkage between Japanese and French or Italians. The link relation in Fig. 6 has been revealed in Fig. 9 by the data analysis. Thus we could clarify the entire picture of the configuration of groups.

Then, we can proceed to a detailed analysis of data without loss of sight of the whole situation. For example, the nations being different in what group of questions and common in what group of questions i.e. simultaneous classification of questions and nations or the universality and particularity of data structure across the nations.

References

The following references are relevant to the various parts of this paper.

Arabie, P., Hubert, L.J. and De Soete, G. (1996) ed.: Clustering and Classification, World Scientific.

Benzécri, J.P. (1973): *L'Analyse des Données*, Dunod.

Benzécri, J.P. (1992) : Correspondence Analysis Hand-Book, Marcel Dekker.

Bock, H.-H. and Polasek, W. (1996) ed.: Data Analysis and Information Systems, Springer.

Borg, I. & Shye, S. (1995): Facet Theory, Form and Content, Advanced Quantitative Techniques in the Social Sciences Series 5, Sage Publication.

Diday, E., J. Lemaire, J. Pouget and F. Testu (1983): *Elements d'Analyse des Données*, Dunod.

Diday, E., G. Celeux, Y. Lechevallier, G. Govaert and H. Ralambondrainy (1989): *Classification automatique et Analyse des Données: Méthodes et environment informatique*, Dunod.

Diday, E. and Y. Leschevallier (1991): Symbolic -Numeric Data Analysis and Learning, -Versailles Sept 91- Nova Science Publisher.

Gaul, W. and Pferfer, D. (1996) ed: From Data to Knowledge, Springer.

Guttman, L. (1994): *Louis Guttman on Theory and Methodology: Selected Writings*, Shlomit Levy ed, Dartmouth.

Hayashi, C. (1956): Theory and example of quantification(II). Proc. Inst. Statist. Math., 3, 69-98.

Hayashi, C. (1974): Minimum dimensional analysis MDA. *Behaviormetrika*, 1, 1-24.

Hayashi, C. (1976): Minimum dimensional analysis MDA-OR and MDA-UO, Essays in Probability and Statistics, Ikeda, S., et al. (eds.), 395-412, Shinko Tsusho Co. Ltd.

Hayashi, C. (1993): Treatise on Behaviormetrics, Asakura Shoten.

Hayashi, C. (1993): Quantification of Qualitative Data --Theory and Method --, Asakura Shoten.

Hayashi, C. (1995): *Changing and Enduring Aspects of Japanese National Character*, The Institute of Social Research, Osaka, Japan.

Hayashi, C. and Suzuki, T. (1986): *Data Analysis in Social Surveys*, Iwanami Shoten. The English version by Hayashi, C. Suzuki,T. and Sasaki, M., "*Data Analysis for Comparative Social Research: International Perspectives*" was published by Elsevier, North-Holland in 1992.

Hayashi, C. and Hayashi, F. (1995): *Comparative Study of National Character*, Proceedings of the Institute of Statistical Mathematics Vol. 43, No.1, 27-80.

Jambu, M. (1989) : *Exploration Informatique et Statistique des Données*, Dunod.

Jambu, M. (1991): Exploratory and Multivariate Data Analysis, Academic Press.

Lebart, L., Morineau, A. and Warwick, K.M. (1984): *Multivariate Descriptive Statistical Analysis*, John Wiley.

Lebart, L. and Salem, A. (1988): *Analyse Statistiques des Données*, Textuelles, Dunod.

Lebart, L. and Salem, A. (1994): Statistique Textuelle, Dunod.

Lebart, L., Morineau, A. and Piron, M. (1995): *Statistique Exploratoire Multidimensionnelle*, Dunod.

Van Cutsem, B. (1994): Classification and Dissimilarity Analysis, Springer.